WLSI-OIAF4HLT 2016

Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies

Proceedings of the Workshop

December 12, 2016
Osaka, Japan

Copyright of each paper stays with the respective authors (or their employers).

ISBN978-4-87974-720-4

Preface

Language technologies and tools (hereafter called language resources) increasingly require sophisticated infrastructures to share, deploy as services, and combine to support research, development, innovation and collaboration. To address this need, several infrastructures have been recently established, including the Language Grid, the Language Application Grid, META-SHARE, DKPro, and CLARIN's Weblicht. While these infrastructures allow users to develop applications using deployed language resources/services, users are typically restricted to tools and resources available through a single infrastructure due to a lack of interoperability. The lack of interoperability among infrastructures has led to duplicated software development efforts as well as redundancy among efforts to manage language resource access, handle licensing concerns, etc.

WLSI-3/OIAF4HLT-2 focuses on the technological and institutional challenges that impact an effort to construct a worldwide interoperable language service infrastructure. It aims to bring together members of the NLP community, including operators of language service infrastructures, but also resource users, developers, and providers, in order to explore and discuss the requirements and desiderata for NLP infrastructures, as well as the opportunities and challenges for enabling interoperable communication among existing infrastructures. The combination of two previously successful workshops addressing NLP infrastructure development and interoperability reflects the recognition of the need for a global effort to achieve mutual interoperability among tools and platforms, and will bring together communities that have previously had little contact.

This volume contains 10 papers from the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI-3/OIAF4HLT-2). The papers are categorized into three parts.

The first part introduces five language service infrastructures that combine natural language processing (NLP) components to develop a language application or analyze text data. Mohanty et al. have proposed Kathaa that enables users to design a language application as an edge-labeled directed acyclic MultiGraph with NLP components. Ide et al. have integrated a workflow system called Galaxy with their Language Application Grid platform to provide users with automated multi-step analyzes and evaluation capabilities. Castilho have analyzed 274 pre-trained models for six NLP tools and reported four potential causes of interoperability problems: encoding, tokenization, normalization, and change over time. Based on these analyses, he has introduced a new tool called Model Investigator to allow model creators to perform automatic sanity checks on their models. To address interoperability problems among language services, Murakami et al. have designed Language Service ontology that enables uses to freely bind language services to a workflow while verifying their composability. Kano have suggested a simplified API for interoperability in order for the developers to more easily employ the functions of Kachako platform.

The second part reports on language service application. Nakaguchi et al. have developed a multi-language support system for international symposiums by combining human inputters and language services on the Language Grid. Assawinjaipetch et al. have proposed complaint classification by employing deep learning techniques with word embedding.

The third part focuses on development of language resources and services, especially low-resourced languages. Aili and Mushajiang have presented how to convert Uyghur dependency Treebank to a universal dependencies version. Luong and Vu have provided a non-expert setup for Kaldi, an open source speech recognition toolkit, to develop a Vietnamese Speech Recognition System. Lastly, to construct annotated corpora, Lu et al. have evaluated an ensemble based pre-annotation approach that combines multiple existing named entity taggers and categorizes annotations into normal ones and candidate ones.

We hope this book will strongly support and encourage researchers who are willing to utilize various

language services worldwide to create customized language applications and multilingual environments. We are grateful to all the participants and those who have supported this workshop.

November 2016

Yohei Murakami (Kyoto University, Japan)
Donghui Lin (Kyoto University, Japan)
Nancy Ide (Vassar College)
James Pustejovsky (Brandeis University)
Workshop Co-Chairs
WLSI-3/OIAF4HLT-2

Organisers

Yohei Murakami (Kyoto University, Japan)
Donghui Lin (Kyoto University, Japan)
Nancy Ide (Vassar College, USA)
James Pustejovsky (Brandeis University, USA)

Programme Committee

Mirna Adriani (University of Indonesia, Indonesia)
Mairehaba Aili (Xinjiang University, China)
Nuria Bel (Universitat Pompeu Fabra, Spain)
Kalina Bontcheva (University of Sheffield, UK)
Nicoletta Calzolari (CNR-ILC, Italy)
Richard Eckart de Castilho (Technische Universität Darmstadt, Germany)
Christopher Cieri (LDC, USA)
Khalid Choukri (ELDA, France)
Riccardo Del Gratta (CNR-ILC, Italy)
Luca Dini (Holmes Semantic Solutions, France)
Zhiqiang Gao (Southeast University, China)
Jens Grivolla (GLiCom, Universitat Pompeu Fabra, Spain)
Hitoshi Isahara (Toyohashi University of Technology, Japan)
Toru Ishida (Kyoto University, Japan)
Yoshinobu Kano (Shizuoka University, Japan)
Monica Monachini (CNR-ILC, Italy)
Weinila Mushajiang (Xinjiang University, China)
Masayuki Otani (Kyoto University, Japan)
Stelios Piperidis (ILSP, Greece)
Vu Hai Quan (University of Natural Sciences, Vietnam National University, Vietnam)
Virach Sornlertlamvanich (SIIT, Thailand)
Andrejs Vasiljevs (Tilde, Latvia)

Table of Contents

Conference Program

Monday, December 12, 2016

9:00–9:10 Opening

9:10–10:40 Language Service Infrastructure 1

9:10–9:40 *Kathaa : NLP Systems as Edge-Labeled Directed Acyclic MultiGraphs*
Sharada Mohanty, Nehal J Wani, Manish Srivastava and Dipti Sharma

9:40–10:10 *LAPPS/Galaxy: Current State and Next Steps*
Nancy Ide, Keith Suderman, Eric Nyberg, James Pustejovsky and Marc Verhagen

10:10–10:40 *Automatic Analysis of Flaws in Pre-Trained NLP Models*
Richard Eckart de Castilho

10:40–11:00 Coffee Break

11:00–12:00 Language Service Application

11:00–11:30 *Combining Human Inputters and Language Services to provide Multi-language support system for International Symposiums*
Takao Nakaguchi, Masayuki Otani, Toshiyuki Takasaki and Toru Ishida

11:30–12:00 *Recurrent Neural Network with Word Embedding for Complaint Classification*
panuwat assawinjaipetch, Kiyoaki Shirai, Virach Sornlertlamvanich and Sanparith Marukata

Monday, December 12, 2016 (continued)

12:00–13:30 Lunch

13:30–15:00 Language Resources and Services

13:30–14:00 *Universal dependencies for Uyghur*
marhaba eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti and Yan Liu

14:00–14:30 *A non-expert Kaldi recipe for Vietnamese Speech Recognition System*
Hieu-Thi Luong and Hai-Quan Vu

14:30–15:00 *Evaluating Ensemble Based Pre-annotation on Named Entity Corpus Construction in English and Chinese*
Tingming Lu, Man Zhu, Zhiqiang Gao and Yaocheng Gui

15:00–15:50 Demo Session and Coffee Break

15:50–16:50 Language Service Infrastructure 2

15:50–16:20 *An Ontology for Language Service Composability*
Yohei Murakami, Takao Nakaguchi, Donghui Lin and Toru Ishida

16:20–16:50 *Between Platform and APIs: Kachako API for Developers*
Yoshinobu Kano

Monday, December 12, 2016 (continued)

16:50–17:00 Closing